Drugi zadatak

```
library(readr)
```

```
dataset <- read_csv("preprocessed_data.csv")
```

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2 Ovisi li veličina podruma o kvartu u gradu?

U podacima postoje varijable TotalBsmtSF te Neighborhood. TotalBsmtSF predstavlja veličinu podruma za pojedinu nekretninu u kvatratnim stopama. Neighborhood predstavlja kvart kojoj nekretnina pripada. Nekretnine s veličinom podruma 0 kvadratnih stopa nećemo koristiti u analizi.

Kako bismo odgovorili na ovo istraživačko pitanje, koristit ćemo metodu ANOVA, odnosno analizirat ćemo varijance. ANOVA je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svake pojedine populacije i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ako postoje razlike u sredinama populacija, one će biti odražene u varijabilnosti među grupama. Analizom varijance htjeli bismo istražutu je li razlika između varijanci slučajna ili nam je statistički značajna.

Budući da u našem pitanju ispitujemo veličinu podruma za različite kvartove, koristit ćemo jedno-faktorski ANOVA model. U jednofaktorskom ANOVA modelu razmatra se utjecaj jednog faktora koji ima k razina.

Neka su:

$$X_{11}, X_{12}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma^2)$$
$$X_{21}, X_{22}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma^2)$$
$$\vdots$$
$$X_{k1}, X_{k2}, \ldots, X_{kn_k} \sim N(\mu_k, \sigma^2)$$

nezavisni uzorci iz k različitih populacija (populacije se razlikuju upravo po razini faktora od interesa). Jednofaktorski ANOVA model glasi:

$$X_{ij} = \mu_i + \epsilon_{ij},$$

gdje je μ_j sredina svake populacije $i = 1, .., k$. Analizom varijance testiramo:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{barem dvije sredine nisu iste.}$$

Jednofaktorski model možemo zapisati i kao

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

gdje je μ srednja vrijednost svih μ_i

$$\mu = \frac{1}{k} \sum_{i=1}^{k} \mu_i,$$

a α_i nazivamo efektom i-tog tretmana. Ekvivalentna hipoteza je sad

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$$
$$H_1 : \text{barem jedna } \alpha_i \text{ je različita od 0.}$$

Razmatramo sljedeće mjere varijabilnosti u podatcima

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})^2 = \text{total sum of squares, ukupna varijabilnost}$$
$$SSA = n\sum_{i=1}^{k}(\bar{x}_{i.} - \bar{x}_{..})^2 = \text{treatment sum of squares, varijabilnost između grupa}$$
$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2 = \text{error sum of squares, varijabilnost unutar grupa}$$

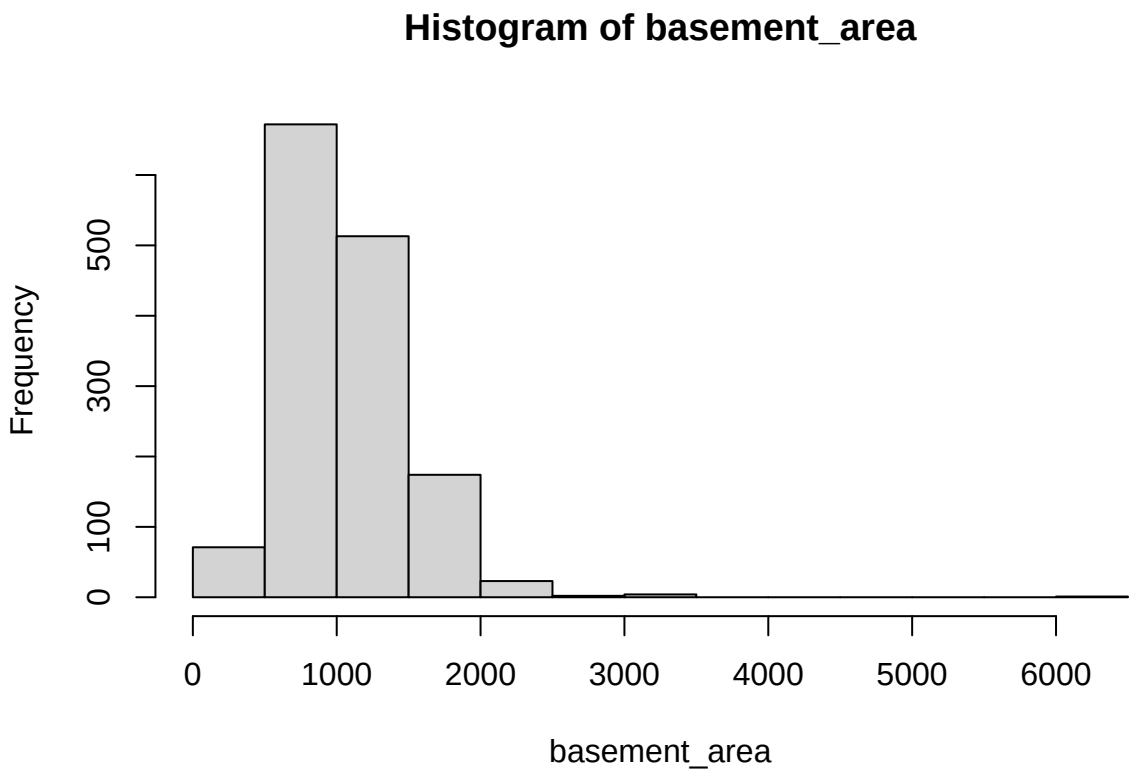
Nadalje, Pretpostavke metode ANOVA su: nezavisnost pojedinih podataka u uzorcima, normalna razdioba podataka i homogenost varijanci među populacijama.

Provjera normalnosti može se za svaku pojedinu grupu napraviti Kolmogorov-Smirnov testom ili Lillieforsovom inačicom navedenog testa. U ovom slučaju razmatrat ćemo veličinu kvarta kao nezavisnu varijablu i veličinu kvarta kao zavisnu varijablu.

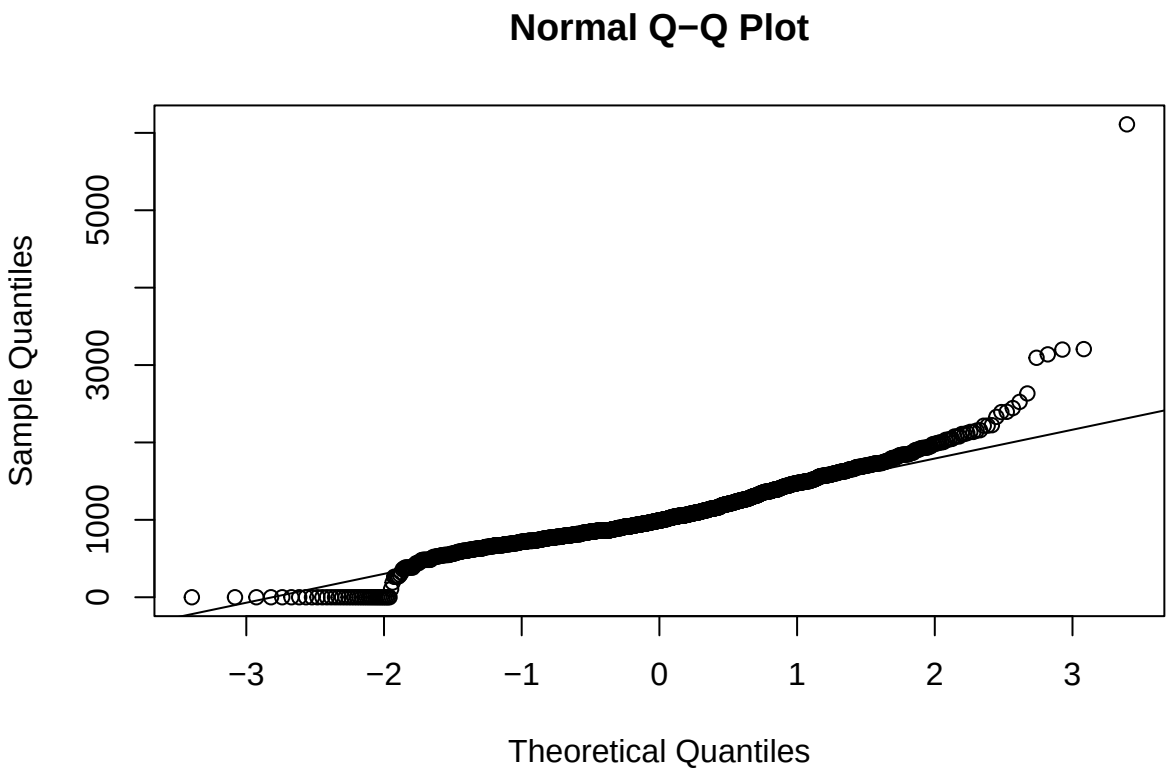
Kako bi provjerili normalnost podataka za veličinu podruma, prvo moramo pripremiti i počistiti podatke

```
# Priprema i čišćenje podataka
basement_area = dataset$TotalBsmtSF

# Grafički prikaz podataka za veličinu podruma
hist(basement_area, breaks = 20)
```




```
qqnorm(basement_area)
qqline(basement_area)
```

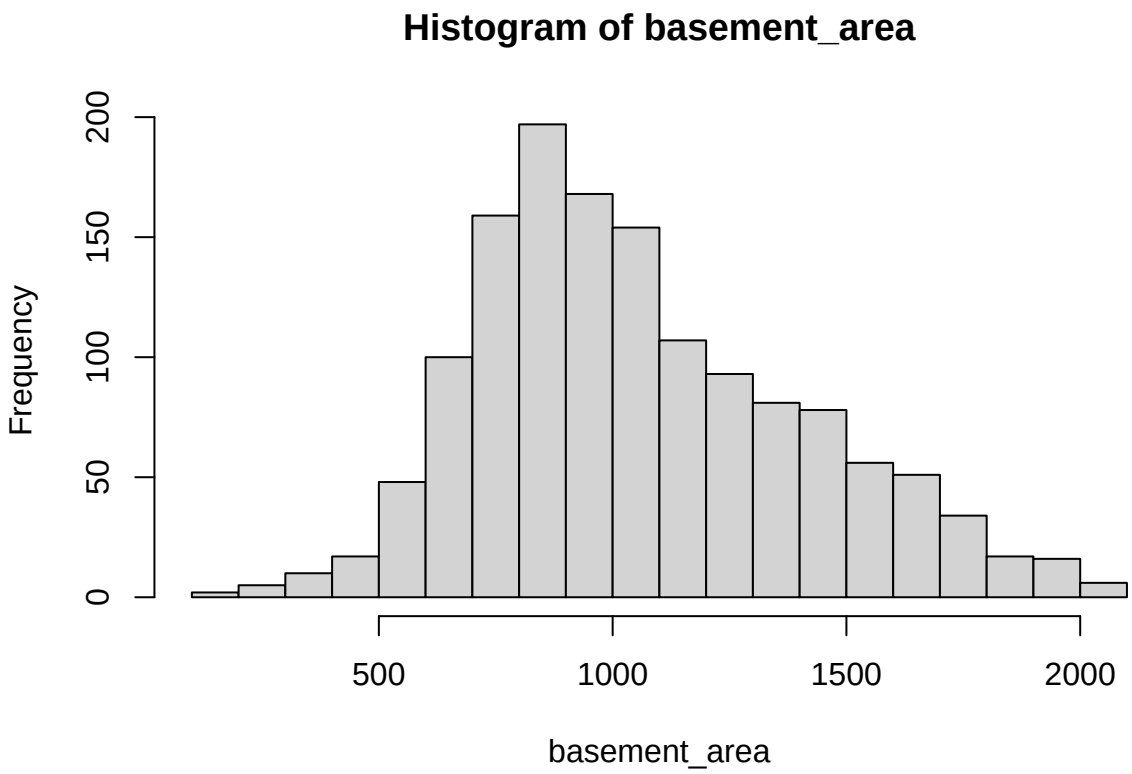


Očito je kako podaci nisu normalni pa ćemo maknuti outliere. Najveći problem predstavljaju podrumi iznad 3000 kvadratnih stopa i oni čija je vrijednost 0.

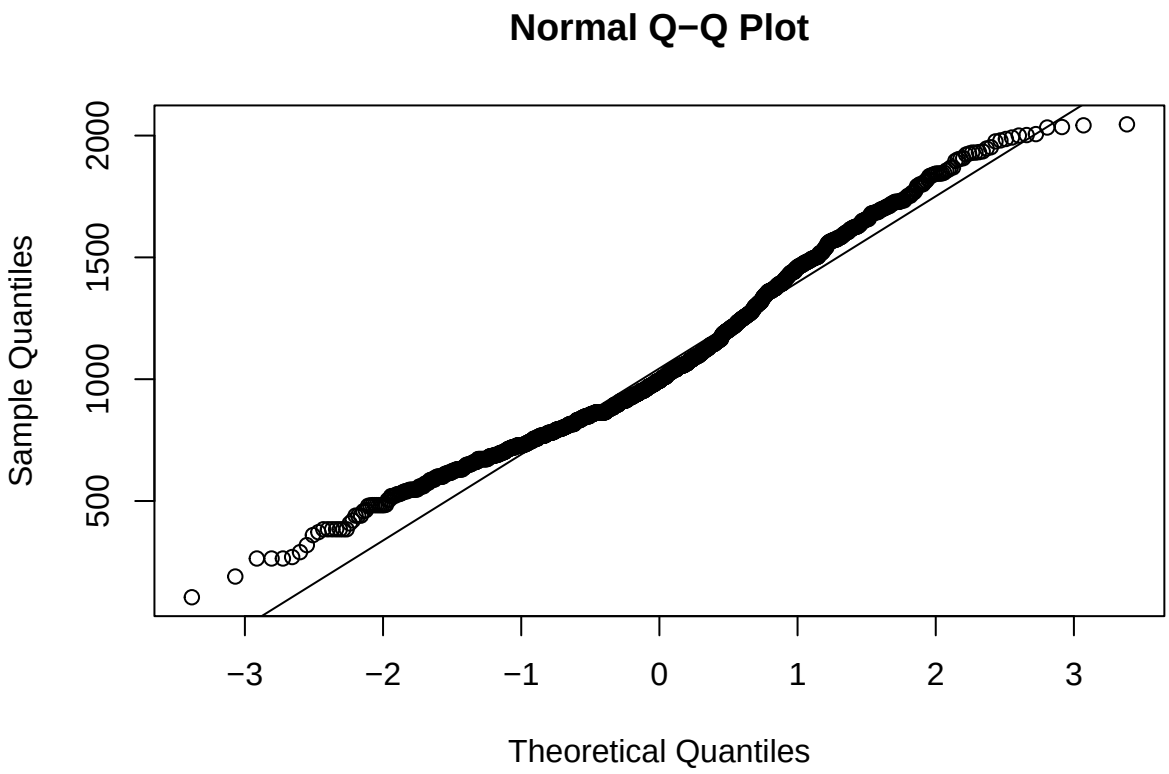
```
# Micanje outliera
quartiles = quantile(basement_area, probs=c(.25, .75), na.rm=FALSE)
IQR = IQR(basement_area)
lower <- quartiles[1] - 1.5*IQR
upper <- quartiles[2] + 1.5*IQR

dataset = subset(dataset, basement_area > lower & basement_area < upper)
basement_area = dataset$TotalBsmtSF

# Grafički prikaz podataka nakon micanja outliera
hist(basement_area, breaks = 20)
```




```
qqnorm(basement_area)
qqline(basement_area)
```



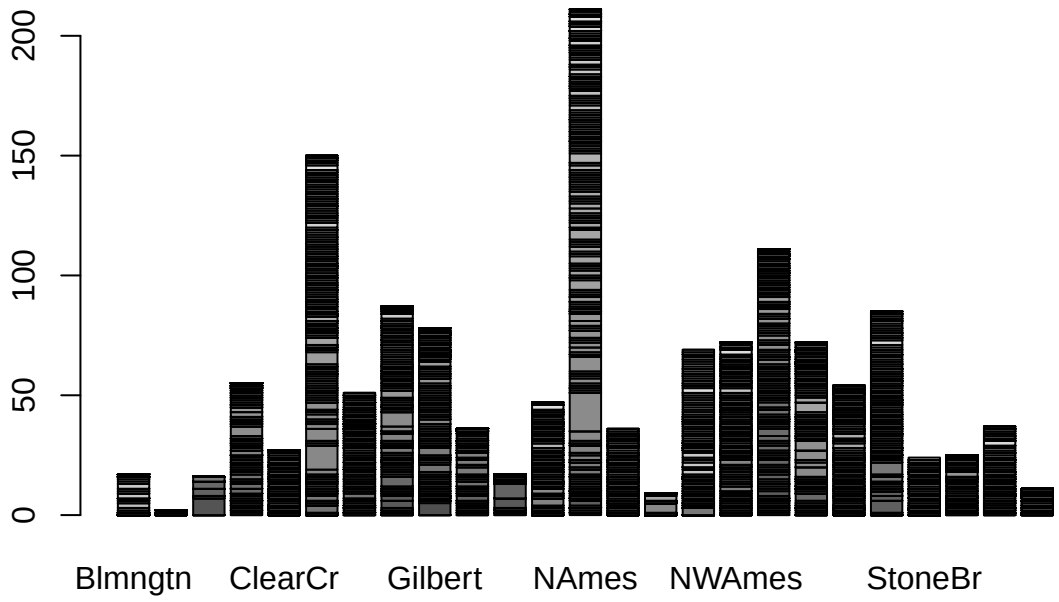
Nadalje, moramo provjeriti ukoliko postoje kvartovi u kojima se nalaze nekretnine koje bi mogle utjecati na daljnju analizu zbog nedovoljne količine podataka

```
neighbourhoods = dataset$Neighborhood
unique_neighbourhoods <- sort(unique(neighbourhoods))

# Broj kvartova u datasetu
length(unique_neighbourhoods)
```

```
## [1] 25
```

```
# Prikaz nekretnina po kvartovima
barplot(table(x= basement_area, y = neighbourhoods))
```




```
counts <- table(neighbourhoods)
counts
```

```
## neighbourhoods
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  IDOTRR
##      17       2      16      55      27     150      51      87      78      36
## MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown  Sawyer SawyerW
##      17      47     211      36       9      69      72     111      72      54
## Somerst StoneBr   SWISU  Timber Veenker
##      85      24      25      37      11
```

Iz daljnje analize maknut ćemo kvartove koji imaju manje od 50 nekretnina kako bi bili sigurniji u normalnost podataka

```
# micanje kvartova koji imaju manje od 50 nekretnina
not_normal_neighbourhoods <- c("Blmngtn", "Blueste", "BrDale", "ClearCr","IDOTRR", "MeadowV","Mitchel",

dataset <- subset(dataset, !Neighborhood %in% not_normal_neighbourhoods)
neighbourhoods = dataset$Neighborhood
length(unique(neighbourhoods))
```

```
## [1] 12
```

Nakon dodatnog čišćenja, preostaje nam 12 kvartove koji su nam važni za nastavak analize. Sada ćemo maknuti outliere u veličini podruma za pojedini kvart.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

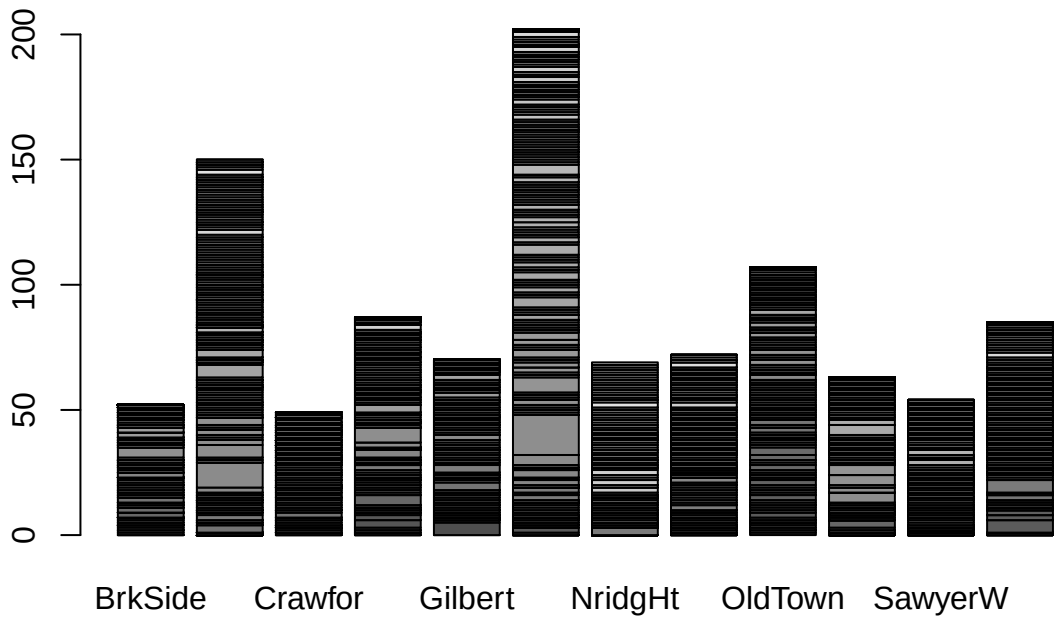
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# cistimo outliere za pojedinačan kvart
dataset <- dataset %>%
  group_by(Neighborhood) %>%
  mutate(Q1 = quantile(TotalBsmtSF, .25),
         Q3 = quantile(TotalBsmtSF, .75),
         IQR = IQR(TotalBsmtSF),
         lower_bound = Q1 - 1.5*IQR,
         upper_bound = Q3 + 1.5*IQR) %>%
  filter(TotalBsmtSF > lower_bound & TotalBsmtSF < upper_bound)
```

```
barplot(table(x= dataset$TotalBsmtSF, y = dataset$Neighborhood))
```



```
sort(unique(dataset$Neighborhood))
```

```
##  [1] "BrkSide" "CollgCr" "Crawfor" "Edwards" "Gilbert" "NAmes"   "NridgHt"
##  [8] "NWAmes"  "OldTown" "Sawyer"  "SawyerW" "Somerst"
```

Nakon čišćenja podataka, možemo provesti Lillieforseov test

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(dataset$TotalBsmtSF)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF
## D = 0.091026, p-value < 2.2e-16
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="BrkSide"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "BrkSide"]
## D = 0.078374, p-value = 0.5901
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="CollgCr"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "CollgCr"]
## D = 0.15775, p-value = 8.8e-10
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="Crawfor"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "Crawfor"]
## D = 0.16077, p-value = 0.002828
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="Edwards"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "Edwards"]
## D = 0.091216, p-value = 0.07084
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="Gilbert"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "Gilbert"]
## D = 0.098398, p-value = 0.08997
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="NAmes"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "NAmes"]
## D = 0.066009, p-value = 0.03227
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="NridgHt"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "NridgHt"]
## D = 0.11189, p-value = 0.03201
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="NWAmes"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "NWAmes"]
## D = 0.14564, p-value = 0.0006527
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="OldTown"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "OldTown"]
## D = 0.085982, p-value = 0.04968
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="Sawyer"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "Sawyer"]
## D = 0.078298, p-value = 0.4397
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="SawyerW"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "SawyerW"]
## D = 0.10272, p-value = 0.1666
```

```
lillie.test(dataset$TotalBsmtSF[dataset$Neighborhood=="Somerst"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$TotalBsmtSF[dataset$Neighborhood == "Somerst"]
## D = 0.11116, p-value = 0.0113
```

Iz rezultata Lillieforsove inačice Kolmogorov-Smirnov testa vidimo da podaci nisu normalne razdiobe. Ukoliko gledamo normalnost veličine podruma unutar kvarta, možemo vidjeti da su unutar šest kvartova od dvanaest podaci normalno distribuirani.

Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

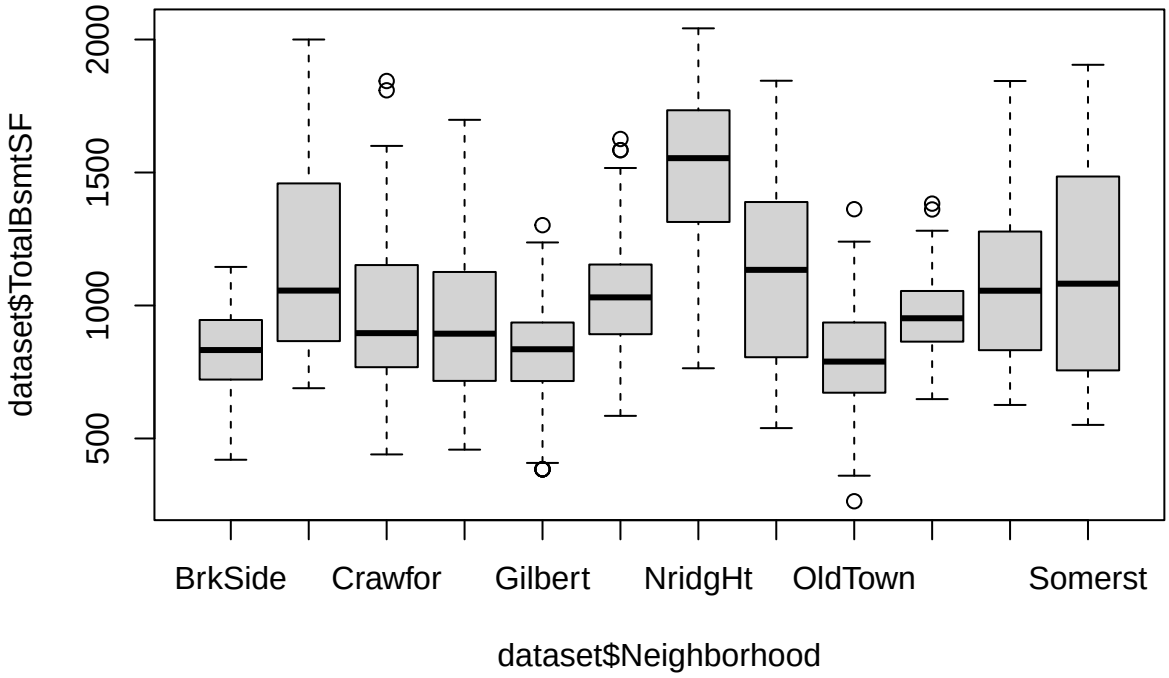
$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \text{barem dvije varijance nisu iste.}$$

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(dataset$TotalBsmtSF ~ dataset$Neighborhood)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  dataset$TotalBsmtSF by dataset$Neighborhood
## Bartlett's K-squared = 146.14, df = 11, p-value < 2.2e-16
```

Bartlettovim testom nezavisnosti dobili smo p-vrijednost iznimno malu što nam sugerira da postoje barem dva kvarta koja imaju različitu varijancu za veličinu podruma. U to se možemo uvjeriti i grafički.

```
# Graficki prikaz podataka
boxplot(dataset$TotalBsmtSF ~ dataset$Neighborhood)
```




```
# Test
a = aov(dataset$TotalBsmtSF ~ dataset$Neighborhood)
summary(a)
```

```
##                          Df   Sum Sq Mean Sq F value Pr(>F)
## dataset$Neighborhood   11 31802055 2891096   39.12 <2e-16 ***
## Residuals            1048 77441129   73894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Grafički prikaz sugerira da postoji jasna razlika u veličinama podruma za pojedinačni kvart, što potvrđuje i ANOVA.

Budući da je p-vrijednost približna nuli, odbacujemo nultu hipotezu u korist alternative, odnosno sa sigurnošću zaključujemo da veličina podruma ovisi o kvartu.